[1] Arnott, S. and Wonacott, A.J., *Polymer* (1966) 7, 157-166.
[2] Smith, P.J.C. and Arnott, S., Acta *Crystallogr*. (1978) A34, 3-11.
[3] Atkins, E.D.T., Nieduszinski, I.A., Mackie, W., Parker, K.D. and Smolko, E.E., *Biopolymers* (1973) 12, 1865-1878.
[4] Atkins, E.D.T., Nieduszinski, I.A., Mackie, W., Parker, K.D. and Smolko, E.E., *Biopolymers* (1973) 12, 1879-1887.
[5] Bian, W., Ph.D. Thesis, Purdue University (2001).
[6] Grant, G.T., Morris, E.R., Rees, D.A., Smith, P.J.C. and Thom, D., *FEBS Letters* (1973) 32, 195-198.
[7] Upstill, C., Atkins, E.D.T. and Attwool, P.T., *Int. J. Biol. Macromol.* (1986) 8, 275-288.
[8] Hamilton, W., *Acta Crystallogr.* (1965) 18, 502-510.
[9] Sheldrick, G.M. and Schneider, T.R., *Methods Enzymol.* (1997) 277, 319-343.
[10] Cruickshank, D.W.J., *International Tables for X-ray Crystallography* (1972) Vol. II, 84-98.

Applications of highly constrained molecular modelling scattering curve fits to biologically important proteins

Stephen J. Perkins

Dept of Biochemistry and Molecular Biology, Royal Free and University Medical School, Rowland Hill Street, London NW3 2PF, UK.

Full molecular structures can be extracted from solution scattering analyses of multidomain or oligomeric proteins if the scattering curve fits are constrained by known small structures for the subunits. All the different possible molecular structures are computed, using as constraints any known covalent connections or symmetry features between the subunits. Each model is assessed for steric overlap, radii of gyration, sedimentation coefficient and R-factor. Filtering leaves a small family of good fit models that corresponds to the molecular structure of interest. These structural analyses often provide new biological insights into function.

Introduction

Solution scattering is a diffraction technique used to study overall structures in solution. A sample is irradiated by a collimated, monochromated beam of X-rays or neutrons. The resulting two-dimensional circularly-symmetric diffraction pattern is recorded on a flat area detector system. Radial averaging leads to a one-dimensional scattering curve. Traditionally these curves leads to structural information at a resolution of about 2-4 nm from calculations of the radius of gyration R_G, the cross-sectional R_G (R_{XS}) and the distance distribution function, and the use of spherical harmonics or genetic algorithms provides an overall view of the macromolecule. This approach provides information on overall macromolecular dimensions and molecular weights from the intensity I(0) at zero scattering angle. In distinction to this traditional approach, the utility of solution scattering has been much improved by means of a novel strategy in which molecular structures are derived directly from the scattering curves. This method starts from known molecular structures for subunits within the macromolecule which are used as tight constraints of the scattering data (reviewed in [1-3]).

In this article, the comparison of a series of investigations [4-18] shows that these studies fall into four distinct types (Table 1). Our most recent studies are discussed to illustrate this approach [14,17].

Scattering modelling potentially provides useful results for a multidomain protein for which an overall structure is unknown, yet molecular structures are available for all the individual domains in it. Such large multidomain proteins are often not crystallisable for reason of interdomain flexibility or surface glycosylation, either of which hinders crystal growth. If crystals are obtained, it is possible that a flexible multidomain arrangement has become frozen into an artefactual snapshot of only one possible conformation. The use of scattering modelling will show what types of domain arrangements are compatible with the solution data. Scattering modelling is also useful in analysing the association of multiple subunits into oligomeric structures. If a biologically important oligomer cannot be crystallised, scattering modelling is a way to obtain a structure. In these applications, the use of known atomic structures as tight constraints to model scattering curves is highly complemented by the continuing growth of the Protein Data Bank, which currently possesses over 14,000 structures (March 2001), as this provides the raw material for these scattering analyses. Indeed the use of these constraints can raise the precision of the scattering models to as high as 0.5 nm.

Scattering modelling is applicable to both X-ray and neutron data. These exhibit very different but complementary properties. X-ray scattering using synchrotron radiation provides high quality curves that are minimally affected by instrumental geometry, as the incident fluxes are sufficiently high to permit the use of ideal pin-hole optics. X-rays visualise the macromolecule in a high positive solute-solvent contrast. More importantly, X-rays also visualise the macromolecule with a hydration shell surrounding it, and this significantly affects the modelling of the scattering curve. In distinction to this, neutron scattering is able to visualise macromolecules in positive and negative contrasts by the use of light and heavy water buffers. The range of scattering densities generated by varying this heavy water content is sufficiently wide to encompass the scattering densities of lipids, protein, carbohydrate and DNA/RNA. Hence, by neutron contrast variation, one component can be matched

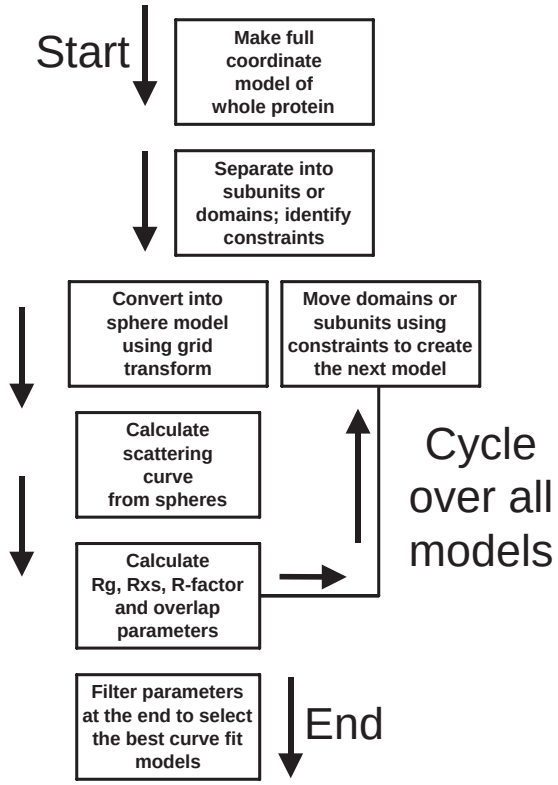

Figure 1: Flow chart of the automated modelling procedure that is used with appropriate modifications for the searches in Table 1.

out by an appropriate choice of heavy water buffer (see below). Neutron scattering is also characterised by the absence of radiation damage effects sometimes encountered with X-ray synchrotron radiation, and also by the ease of determining molecular weights. It is also useful that to a good approximation the hydration shell is not visible in neutron scattering, and this simplifies the application of modelling strategies.

Method for curve fits

The first stage in constrained scattering modelling is of course the experimental data acquisition itself. Our data have been obtained at the SRS Daresbury (X-rays), ISIS Rutherford-Appleton (neutrons) and the Institut Laue Langevin (neutrons) laboratories. Technical details are given in [3]. Analyses to obtain the RG and I(0) values from Guinier plots and distant distribution functions are completed in full before any modelling is initiated. The second stage is to identify the atomic structures that best represent the subunits of the full structure to be modelled, whether these be directly obtained from crystal or NMR structures, or indirectly by the application of homology (or "comparative") modelling. The third stage is to model the X-ray and neutron scattering curves I(Q) using small spheres of uniform density

to represent the protein structure. Their total volume must be the same as that of the dry protein. Curves are calculated from Debye's Law adapted to small spheres [1,2]. These spheres have to be sufficiently small (about 0.6 nm) so that their form factor in the Debye equation is almost invariant in the Q scattering range used (Q = 4B sin 2 /8; 22 = scattering angle; 8 = wavelength).

The first of the four types of constrained scattering modelling listed in Table 1 are the calibration studies required to establish the technique [4-6]. For these, the crystal structures for ∀-chymotrypsin, ∃-trypsin, $\forall_1$-antitrypsin and pentameric serum amyloid P component (SAP) correspond to structures that are rigid and well-defined in solution and cover a wide molecular weight range of 23,200-127,000. These calibrations showed that the same single density approach under conditions of high solute-solvent contrasts worked well for both proteins and glycoproteins. Similar good fits were obtained with X-ray data in high positive contrasts and 100% heavy water neutron data in high negative contrasts, even with carbohydrate contents as high as 50% as found in carcinoembryonic antigen (CEA) [16]. The calibrations also showed that the hydration shell could be neglected in the neutron fits, but is required in the X-ray fits. This shell corresponds to a water monolayer surrounding the protein surface and is well-modelled by 0.3 g of water/g glycoprotein and an electrostricted volume of 0.0245 nm^3 per bound water molecule. Hydration shells are best modelled by adding spheres in a uniform layer to the surface of the model that is adjusted to reach the required hydrated volume [6]. The calibrations also show that no instrumental corrections for X-ray wavelength spread or beam divergence are required, although these are necessary for neutron cameras for reason of their larger physical dimensions. Neutron curve fits sometimes deteriorate at large Q. This is usually attributable to a small residual flat background that arises from incoherent scatter from protons in the sample.

The remaining three types of constrained modelling are automated conformational searches (Table 1). All three utilise the procedure summarised in Figure 1. Script files written for standard molecular graphics software packages generate a full range of models starting from atomic structures for the individual domains or subunits. Each one is readily converted into a sphere model by a grid transformation, from which the scattering curve is calculated for comparison with experimental data.

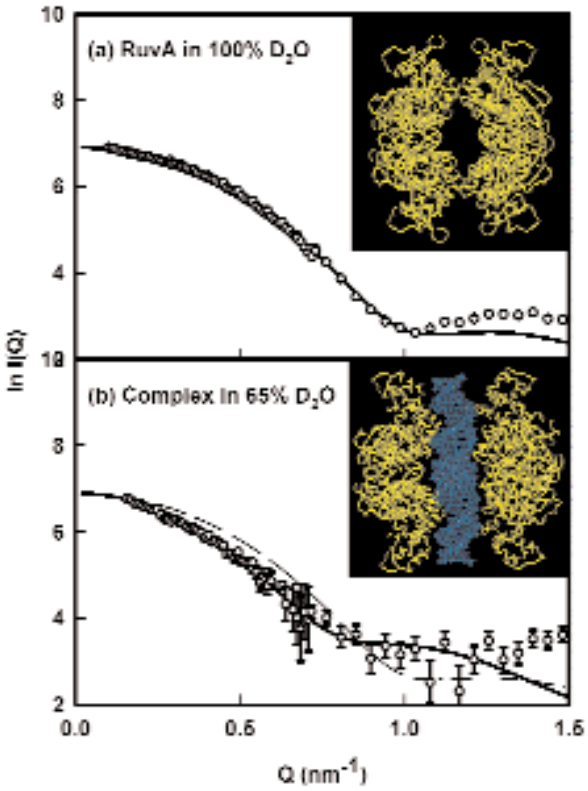

Figure 2: Neutron scattering curve fits for (a) the RuvA octamer (yellow: seen edge-on) and (b) its complex with the four-way Holliday junction DNA (blue). The curve fits in 100% D_2O for RuvA and in 65% D_2O for the complex (DNA invisible) are shown by the continuous lines. The dashed line in (b) corresponds to the unbound RuvA octamer and illustrates the large structural change on complex formation. Adapted from [14].

Up to four filters are used to identify the best models and to remove unsatisfactory models: (i) The systematic creation of models can result in physically unreasonable steric overlap between domains or subunits. This is readily detected by the grid transformation which will give too few spheres if significant overlap occurs, so there was a requirement for at least 95% of the expected total of spheres in a good model. (ii) The R_G and R_{XS} values were determined from the calculated curves in the same Q ranges used in the experimental Guinier fits of the experimental data. These are required to be within 5% of the experimental values in a good model. (iii) Sedimentation coefficients were calculated from the hydrated sphere models used for X-ray fits for comparison with the experimental values as an independent monitor of the search. (iv) The remaining good models were then assessed using a goodness-of-fit R-factor computed by analogy with the same term used by crystallographers. This rank ordering of all the good

Protein[1]	Molecular weight	Scattering data[2] method	Search	Number of models tested	Constraints	Reference
(a) Calibration Studies						
$\forall_1$-antitrypsin	51,500	X, N, S	Manual	1	Crystal structure	[4]
Trypsin/chymotrypsin (bovine)	23,200-25,600	X, N, S	Manual	4	4 crystal structures	[5]
SAP pentamer	127,000	X, N	Manual	3	Pentamer structure	[6]
(b) Translational and/or rotational searches of separate subunits						
IgM and its fragments	976,000	X, S	Manual	~200	2 Fab and 1 Fc structures	[7]
IgG1 and IgG2 (bovine)	144,000	N	Semi-automated	~200	Crystal structure	[8]
IgE-Fc	75,300	X, N	Automated	37,440	2 Ig folds and 1 Fc structures	[9]
Carcinoembryonic antigen	152,500	X, N, S	Automated	20,280	1 V-type and 6 C2-type Ig folds	[10]
Factor VIIa	51,400	X, N, S	Automated	15,625	4 FVIIa domains	[11]
Tissue factor-FVIIa complex	76,200	X, N, S	Automated	37,044	2 FVIIa and TF structures	[11]
Factor 1	85,300	X, N	Automated	9,600	5 factor I domains in 2 lobes	[12]
(c) Symmetry-constrained translational searches						
AmiC trimers (*P. aeruginosa*)	127,900	X, N	Automated	21	Monomer structure; trimeric symmetry	[13]
SAP decamer	254,000	X, N	Automated	640	Pentamer structure; axial symmetry	[6]
RuvA (*M. leprae*)	165,700	N	Automated	120	E.Coli RuvA structure; axial symmetry	[14]
RuvA-Holliday junction complex	205,100	N	Automated	200	E.Coli RuvA structure; axial symmetry	[14]
MFE-23 (scFv antibody: *E. Coli.)*	27,200	N, S	Manual	3	MFE-23 structure in crystal lattice	[15]
(d) Covalently-connected domain searches by molecular dynamics						
Carcinoembryonic antigen	152,500	X, N, S	Manual	1	1 V-type,3 I-type,3C2-type; 6 CD2 linkers	[16]
IgA1	164,000	X, N	Automated	12,000	2 Fab and 1 Fc structures; 2 covalent linkers	[17]
Factor H	150,000	X, N, S	Automated	16,752	20 SCR structures; 19 covalent linkers	[18]

Table 1: Summary of constrained scattering curve modelling analyses

[1] Proteins are of human origin unless specified otherwise.

[2]. X, X-ray scattering; N, neutron scattering; S, sedimentation coefficients.

models defines the best-fit structures, which are then examined in more detail.

Translational and/or rotational searches of separate domains

Translational and rotational searches of domain fragments provides a straightforward approach for obtaining curve fits. Our first attempt to model a multidomain protein in terms of component crystal structures was made with the antibody IgM. This is a pentameric molecule in which 10 four-domain Fab and 5 four-domain Fc fragments and 11 other domains form a large planar structure in solution. This total of 71 domains was analysed in a stepwise fashion, in which scattering curves for the four-domain Fab fragment, the ten-domain Fab_2 fragment, and the 21-domain Fc_5 fragment were each individually modelled. The scattering curve for the intact IgM structure was then modelled in terms of these three models. The curve fits were achieved using manual rotational and translational searches of small sphere models. The outcome of the modelling was illuminating, as the homology model of the Fc_5 fragment revealed the likely position of a binding site for C1q of complement, which was masked in free IgM, but is exposed when the five Fab_2 fragments are bound to an activating surface [7].

These simple searches were automated for the modelling of the two Fab and one Fc fragments connected by hinge peptides in intact bovine IgG1 and IgG2 antibodies. A translational search optimised the relative positions of these three individual fragments within each molecule. The best-fit models showed that both IgG molecules possessed extended arrangements of these three fragments that allowed full access to the hinge peptides and receptor sites at the centre of each structure [8]. Automation was also applied to the study of the six-domain Fc fragment of human IgE, in which there is an additional pair of domains $(C\gamma2)_2$ in place of the hinge peptide in IgG. This modelling in terms of small sphere models proved to be more complex than anticipated, and was completed by allowing for 37,000 rotations and translations of the $(C\gamma2)_2$ domains and possible rotations in the other domains. The best fit model was defined as the mean structure of the 100 models with the best R-factor values. This showed that a bent IgE-Fc model with a $(C\gamma2)_2$ rotation of 70° gave a very good X-ray curve fit, and accounted for reports that the intact IgE antibody is itself bent [9]. Another translational-rotational search of the five

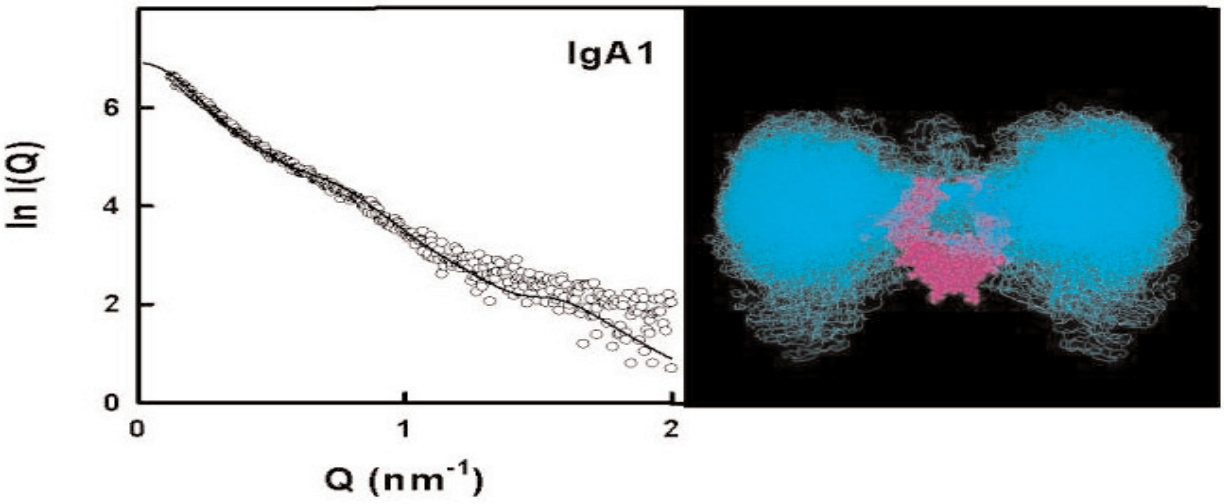

Figure 3: X-ray scattering curve fit for IgA1. The superimposition of 104 best-fit IgA1 models is shown, with the Fc fragment depicted in purple at the centre, and the 104 pairs of Fab fragments are shown in cyan. Adapted from [17].

domains in factor I of complement showed that a two-lobed structure provided a plausible explanation for why its five domains did not form an extended structure in solution [12].

Curve fits can also be obtained by rotation searches. CEA contains seven Ig fold domains, and is 50% carbohydrate. Here, scattering modelling was based on an automated set of systematic domain rotations based on a fixed interdomain separation. The 100 best-fit CEA models showed an extended zig-zag structure with carbohydrate chains extended away from its surface [10].When the crystal structure of an anti-CEA antibody MFE-23 became available, the remodelling of the entire zig-zag CEA structure in terms of the linker conformation observed in the CD2 crystal structure showed that it was possible both to improve the curve fit and to propose a model for its complex with MFE-23 [16]. Another automated domain rotation search showed that factor VIIa of blood coagulation had an extended four-domain conformation [11].

Symmetry-constrained subunit translational searches

Proteins often form oligomeric structures that can be analysed by scattering, for which the most notable feature is that symmetry considerations play a role in the curve modelling. AmiC is a two-domain periplasmic binding protein that was shown by scattering to exist in a monomer-trimer equilibrium. This was unexpected as its crystal structure revealed an antiparallel dimer. As a trimeric structure has a three-fold axis of symmetry, the scattering modelling of the trimer was achieved by arranging the long axes of three monomers parallel to each other about this axis, and optimising the position of the monomers using 21 translations [13]. Asymmetric associations of AmiC monomers and dimers are ruled out as valid scattering models, since these would form indefinitely self-associated structures, which are not observed. A variant of this type of modelling is to use observed crystal lattice packing arrangements as scattering models for oligomer structures in solution, which is possible if symmetric subunit associations are observed [15].

Symmetry considerations lead to the most precise scattering modelling fits. Serum amyloid P component (SAP) is a disc-like pentameric molecule that forms very stable face-to-face decamers in the absence of calcium [6]. RuvA is another disc-like tetrameric molecule that forms face-to-face octamers and binds to a four-stranded DNA helix structure called a Holliday junction [14]. Both SAP and RuvA were modelled by taking advantage of the five-fold or four-fold symmetry axis perpendicular to the plane of SAP or RuvA respectively. In these searches, one molecule was held fixed while the other was translated along the symmetry axis in 0.1 nm steps. Both searches gave two best-fit minima that correspond to the two possible structures formed by face-to-face contacts (cf: Figure 2). For SAP, one decamer structure was rejected as it is devoid of the Trp residues known to be present at the interface between the two pentamers. However there was uncertainty in the outcome of the SAP modelling as the relative rotation between the two pentamers could not be well determined. For RuvA, one of the two octamer structures was rejected as it is devoid of charged groups that would readily account for its formation. This time, there is no uncertainty in the relative rotation between the two tetramers as only one aligned orientation permits the formation of a clear DNA-binding groove between them. This modelling was extended to include the use of neutron contrast variation, which enabled the DNA in RuvA complexed with a Holliday junction to be masked out in the scattering experiment. Neutron data recorded using 65% D_2O permitted the modelling of the protein only in the complex to be carried out (Figure 2). This revealed a gap between the two tetramers, which corresponded in size to the width of a DNA double helix. The modelling concluded by showing that this octameric complex existed in solution, and clarified the significance of crystallographic studies on the tetrameric and octameric forms of this complex.

Covalently-connected domain searches by molecular dynamics

To model a multidomain protein, it is sometimes better to replace the straightforward rotational and translational searches using a fragmented subunit structure with a full covalently-connected structure. This avoids the arbitrary nature of these searches as each model is now stereochemically correct before it enters the curve fit process. If the conformationally variable linker peptides between domains are modelled using structural libraries calculated by molecular dynamics simulations, full models of the protein are generated by an automated process that assembles randomly selected linker peptides with the individual domains to create the full structure. Curve

fits then become a trial-and-error procedure that is left to run until a sufficient number of good-fit solutions are obtained. This was applied to determine the solution structure of the antibody IgA1 as an assembly of three well-defined rigid homology models for the two Fab and one Fc fragments joined by two extended 23-residue glycosylated hinge peptides whose conformations were unknown [17]. The best-fit structures showed that the hinge peptides were highly extended and positioned the two Fab fragments far away from the Fc fragment in IgA1 (Figure 3). Such a structure accounted for the location of the IgA receptor site at the centre of the Fc fragment, rather than at the top of the Fc fragment as found in IgG antibodies, as the position of the IgA1 Fab fragments do not obstruct access to this site. The importance of this alternative modelling approach for IgA1 is that the use of translations and rotations became too complex to interpret in this particular case, and gave inconsistent results [17]. The use of molecular dynamics to create peptide libraries offers a powerful strategy that is applicable to other multidomain proteins. For example, its use with the 20 SCR domains and 19 linker peptides in factor H of complement demonstrated that this protein possessed a folded-back domain structure in solution [18].

Conclusions

What is common to all three methods of constrained scattering fits is that, if relevant small crystal structures are available, a large number of possible macromolecular models are generated. Systematic comparisons with the experimental scattering curve then leaves a small number of best-fit structures. The biological significance of these studies is similar to that of standard protein homology modelling, in that the three-dimensional proximity arrangement of key amino acid residues are indicated. For example, the antibody analyses indicated the accessibility of several key residues relative to the other domains for interactions with their receptors. While a good curve fit is only a test of consistency, and will not constitute a unique structure determination, the modelling becomes more useful when the constraints become stronger. Frequently, the modelling complements newly-determined crystal structures by clarifying important details in these. Alternatively, it is possible to analyse structures that crystallographers are unable to solve.

Acknowledgements

The Wellcome Trust, the Biotechnology and Biological Sciences Research Council, and the Clement Wheeler-Bennett Trust are thanked for grant support. Many biochemical collaborators are thanked for their generous provision of samples, and instrument scientists at the SRS, ISIS and ILL provided invaluable support.

References

[1] Perkins, S. J., Ashton, A. W., Boehm, M. K., Chamberlain, D. (1998) *Int. J. Biolog. Macromo*l. 22, 1-16.
[2] Perkins, S. J., Ullman, C. G., Brissett, N. C., Chamberlain, D., Boehm, M. K. (1998) *Immunol. Reviews*, 163, 237-250.
[3] Perkins, S. J. (2000) In *Protein-Ligand Interactions: A Practical Approach* (Eds. B. Chowdhry and S. E. Harding) 1, 223-262.
[4] Smith, K. F., Harrison, R. A. & Perkins, S. J. (1990) *Biochem. J.* 267, 203-212.
[5] Perkins, S. J., Smith, K. F., Kilpatrick, J. M., Volanakis, J. E., Sim, R. B. (1993) *Biochem. J.* 295, 87-99.
[6] Ashton, A. W., Boehm, M. K., Gallimore, J. R., Pepys, M. B., Perkins, S. J. (1997) *J. Mol. Biol.* 272, 408-422.
[7] Perkins, S. J., Nealis, A. S., Sutton, B. J., Feinstein, A. (1991) *J. Mol. Biol.* 221, 1345-1366.
[8] Mayans, M. O., Coadwell, W. J., Beale, D., Symons, D. B. A., Perkins, S. J. (1995) *Biochem. J.* 311, 283-291.
[9] Beavil, A. J., Young, R. J., Sutton, B. J., Perkins, S. J. (1995) *Biochemistry*, 34, 14449-14461.
[10] Boehm, M. K., Mayans, M. O., Thornton, J. D., Begent, R. H. J., Keep, P. A., Perkins, S. J. (1996) *J. Mol. Biol.* 259, 718-736.
[11] Ashton, A. W., Boehm, M. K., Johnson, D. J. D., Kemball-Cook, G., Perkins, S. J. (1998) Biochemistry, 37, 8208-8217.
[12] Chamberlain, D., Ullman, C. G., Perkins, S. J. (1998) *Biochemistry*, 37, 13918-13929.
[13] Chamberlain, D., O'Hara, B. P., Wilson, S. A., Pearl, L. H., Perkins, S. J. (1997) *Biochemistry*, 36, 8020-8029.
[14] Chamberlain, D., Keeley, A., Aslam, M., Arenas-Licea, J., Brown, T., Tsaneva, I. R., Perkins, S. J. (1998) *J. Mol. Biol.* 284, 385- 400.
[15] Lee, Y.-C., Boehm, M. K., Perkins, S. J. (2001) In preparation.
[16] Boehm, M. K., Perkins, S. J. (2000) *FEBS*

Letters 475, 11-16.
[17] Boehm, M. K., Woof, J. M., Kerr, M. A., Perkins, S. J. (1999) *J. Mol. Biol*. 286, 1421-1447.
[18] Aslam, M., Perkins, S. J. (2001) Submitted.

Fibre Diffraction Using the BioCAT Facility at the Advanced Photon Source

T.C. Irving* and R. F. Fischetti.

The Biophysics Collaborative Access Team (BioCAT), Dept of Biological Chemical, and Physical Sciences, Illinois Institute Of Technology, Chicago, IL, 60616 USA.

***Corresponding author: T.C. Irving (312) 567-3489 Fax: (312) 567-3494 e-mail: irving@biocat1.iit.edu**

The BioCAT undulator-based beamline at the Advanced Photon Source (APS), Argonne IL, USA is a state-of-the-art instrument for biological non-crystalline diffraction and X-ray absorption spectroscopy that is generally available to the international scientific community. The design features of this instrument and the unique source properties of the APS allow collection of fibre diffraction patterns of exceptional quality from complex, weakly diffracting biological systems. The small focal spots achievable with this instrument (~40 x 200 microns) has allowed excellent discrimination of fine detail in fibre patterns from muscle and connective tissue as well as detection of weak diffraction features in the presence of large backgrounds. The high X-ray flux of the instrument (~1.5 x 10^{13} photons/s at 12 keV) permits dynamical experiments on these systems with very fast time resolution.

Introduction

The Biophysics Collaborative Access Team (BioCAT) is a US National Institutes of Health - Supported Research Center dedicated to structural studies of partially ordered biological materials using small-angle X-ray diffraction (SAXS), small-angle solution scattering (SAS), and x-ray absorption (XAS) spectroscopy at the Advanced Photon Source (APS) Argonne National Labs, Argonne, IL. The BioCAT facility is open to all researchers on the basis of peer reviewed beam time proposals. Central to the facility is an undulator-based beamline located on Sector 18 at the APS. First monochromatic light from this instrument was achieved in September of 1997. Since then, in collaboration with a number of external groups, we have collected a considerable body of data, most of it in the area of small-angle fibre diffraction. Here we present some of these diffraction images obtained from a variety of complex biological tissues that demonstrate the